

**Robert Jay Milewski E-mail:**  
milewski@cedar.buffalo.edu ·  
**Venu Govindaraju E-mail:**  
govind@cedar.buffalo.edu ·  
**Anurag Bhardwaj E-mail:**  
ab94@cedar.buffalo.edu

# Automatic Recognition of Handwritten Medical Forms for Search Engines

the date of receipt and acceptance should be inserted later

**Abstract** A new paradigm, which models the relationships between handwriting and topic categories, in the context of medical forms, is presented. The ultimate goals are (i) a robust method which categorizes medical forms into specified categories, and (ii) the use of such information for practical applications such as an improved recognition of medical handwriting or retrieval of medical forms as in a search engine. Medical forms have diverse, complex and large lexicons consisting of English, Medical and Pharmacology corpus. Our technique shows that a few recognized characters, returned by handwriting recognition, can be used to construct a linguistic model capable of representing a medical topic category. This allows (i) a reduced lexicon to be constructed, thereby improving handwriting recognition performance, and (ii) PCR (Pre-hospital Care Report) forms to be tagged with a topic category and subsequently searched by information retrieval systems. We present an improvement of over 7% in raw recognition rate and a mean average precision of 0.28 over a set of 1175 queries on a data set of unconstrained handwritten medical forms filled in emergency environments.

**Keywords** Handwriting Analysis, Language Models, Pattern Matching, Retrieval Models, Search Process

## 1 Introduction

This paper describes the first automatic recognition system for handwritten medical forms. In the United States, any pre-hospital emergency medical

---

Center of Excellence for Document Analysis and Recognition  
UB Commons, 520 Lee Entrance, Suite 202  
Amherst NY 14228

<sup>0</sup> This work was supported by the National Science Foundation.

- 
- 1) Form, agency, and ambulance vehicle Identification
  - 2) Patient and physician contact information
  - 3) Care in progress on arrival and mechanism of injury
  - 4) Dispatch Information
  - 5) Patient Transfer Information
  - 6) Rescue times between rescue and transport phases
  - 7) Extrication and patient vehicle information
  - 8) Chief Complain
  - 9) Subjective Assessment
  - 10) Presenting Problem
  - 11) Past Medical History
  - 12) Vital/Signs
  - 13) Objective Physical Assessment
  - 14) Physical Assessment Extension and/or Comments
  - 15) Treatment Given
  - 16) Ambulance Crew Identification
- 

**Table 1** PCR Categories

care provided must be documented. Departments of Health for each state provide a standard medical form to document all patient care from the beginning of the rescue effort until the patient is transported to the hospital. State laws require that emergency personnel fill out one form for each patient. Automatic recognition and retrieval of these forms is quite challenging for several reasons: (i) Handwritten data in the form is unconstrained in terms of writing styles, variability in font type or size and choice of text due to emergency situations, (ii) Form images are noisy since they are obtained from carbon copies of the original forms, (iii) Dictionary of medical words is huge with over 40,000 words which leads to poor recognition results.

Figure 1 shows an example Pre-Hospital Care Report (PCR) [67] form which contains 16 information regions (see Table 1). Handwriting, from PCR regions 8, 9, 11, 13 and 14 are used for recognition and retrieval analysis. There are two phases to our research: (i) the recognition of handwriting on the medical form, and (ii) a medical form query retrieval engine. Handwriting recognition is used to tag medical forms with a topic category to subsequently improve recognition performance. The medical forms reflect large lexicons containing Medical, Pharmacology and English corpus. While current state of the art recognizers report recognition performance between  $\sim 58-78\%$ , on comparable lexicon sizes in the postal application [36] [68] [69], our experiments show  $\sim 25\%$  raw match recognition performance on the medical forms. This underscores the extremely complicated nature of medical handwriting (Figure 1). We have developed a method of automatically determining the topic category of a PCR form using machine learning and computational linguistics techniques. We demonstrate the strategy for improving the raw word recognition rate by about 7% for a lexicon size of over 5,000 words.

## 2 Background

Though the task of efficient retrieval of text documents has been addressed by information retrieval community for several years [70], robust document

[illegible]

**Fig. 1** Pre-Hospital Care Report (PCR) Labeled [67]

search and retrieval has received some considerable attention lately [16]. The existing methods for document retrieval can be broadly classified into two categories - (i) OCR based methods [28] [58] [65] and (ii) Word image matching based methods [56] [54] [2] [55] [64]. On one hand word image matching based methods rely heavily on the proper selection of image features [53] and similarity methods [55] [2], the OCR based methods depend on the word recognition accuracy. It has been shown that higher word recognition error rate adversely affects the document retrieval performance [40] [14]. Therefore, an improved word recognition algorithm forms a basis for an efficient document retrieval system.

The basis for reducing the lexicon to improve recognition is a well researched strategy in handwriting recognition [26] [68]. Although handwriting

recognition and lexicon pruning/reduction [43] have been researched substantially over the years, many challenges still persist in the offline domain. Word recognition applications range from automated check recognition [35], postal recognition [20], historical documents recognition [21] [25] and now emergency medical documents [45] [46] [47]. Strategic recognition techniques for handwriting algorithms such as Hidden Markov Models (HMM) [37] [44] [48] [31] [11] [18] [17], Artificial Neural Networks (ANN) [50] [6] [13] [22] [12], Boosted Decision Trees [30] and Support Vector Machines (SVM) [1] [7] have been developed. Lexicon reduction has been shown to be critical to improvement of performance primarily because of the minimization of possible choices [26]. Even the systems dealing with a large vocabulary corpus have been successful [37] [38] [72].

Lexicon reduction schemes in general, rely upon finding a specific topic of the document and then using a fixed smaller vocabulary of the chosen category as the reduced lexicon. This is usually achieved by performing categorisation of the OCR'd document text which is noisy. Bayer et al. [3] in their work learn the noise model of the OCR using word substrings extracted with an iterative procedure. Taghva et al. [63] study the performance of a naive bayes classifier over 400 documents recognized with an OCR at a word error rate of nearly 14%. 6 categories out of 52 are analyzed and the highest rate of correct classification achieved is 83.3%. However, both these strategies have been applied to machine print OCR'd text where the noise level is not as high as the handwritten documents. In the context of medical forms, where the word recognition rate is very low ( $\sim 25\%$ ) and only few characters are recognized with high confidence scores, such methods are not applicable. Vinciarelli et al. [66] study noisy text categorization over synthetic handwritten data. In this research, noisy data is obtained by changing certain percentage of characters obtained from the OCR. However this method only handles the case when the character is changed to another list of known characters, whereas in the text obtained from medical forms, there are slots for potentially unknown or human unreadable characters.

Additionally, some lexicon reduction strategies have used the extraction of character information for lexicon reduction, such as that by Guillevic, et al. [27]. However, such strategies reduce the lexicon for a single homogeneous category, namely cities within the country of Finland. In addition, usage of word length estimates for a smaller lexicon are available [27]. Caesar, et al. [8] also state that prior reduction techniques [60] [61] [51] are unsuitable since they can only operate on very small lexicons due to enormous computational burdens [8]. Caesar [8] further indicates that Suen's [62] approach of n-gram combinatorics is sensitive to segmentation issues, a common problem with medical form handwriting [8]. However, Caesar's method [8] and those which are dependent on using the character information, and/or the character information of only one word to directly reduce the lexicon, suffer if one of the characters is selected incorrectly [8]. This is observable in the cursive or mixed-cursive handwriting types.

Many existing schemes, such as that of Zimmermann [71], assume that some characters can be extracted. However, in the medical handwriting domain this task is error prone. Therefore, operating a reduction scheme which

can be robust to incorrectly chosen characters is necessary. We use sequences of characters to determine the medical topic category which has a lexicon of its own, thereby reducing the issues of using the character information directly. Similar to the study by Zimmermann et al. [71], the length of words are used with phrases.

Kaufmann, et al. [34] present another HMM strategy which is primarily a distance-based method and uses model assumptions which are not applicable in the medical environment. For example, Kaufmann [34] assumes that “...people generally write more cooperatively at the beginning of the word, while the variability increases in the middle of the word.” In the medical environment, variability is apparent when multiple health care professionals enter data on the same form. The medical environment also has exaggerated and/or extremely compressed word lengths due to erratic movement in a vehicle and limited paper space. Kaufmann [34] only provides a reduction of 25% of the lexicon size with little to no improvement in error rate, and the experiments are run only on a small sample of words.

### 3 Lexicon Reduction

This research proposes the following hypothesis which is verified experimentally: A sequence of confidently recognized characters, extracted from an image of handwritten medical text, can be used to represent a topic category. The construction of medical form training and test set has been created manually. A software data entry system has been developed which allows human truthers to segment all PCR form regions and words, and provide a human interpretation for the word, denoted as the truth. Truthing is done in two phases: (i) the digital transcription of medical form text, and (ii) the classification of forms into topic categories. The distribution of PCR forms under each category is approximately equal in both the training and test set (see Table 2). The task has been supervised and performed by a health care professional with several years of field emergency medical services (EMS) experience. This emergency medical data set is the first of its kind.

A PCR can be tagged with multiple categories. In our data set, no form had more than five category tags. The subjectivity involved in determining the categories makes the construction of a hierarchical chart representing all patient scenarios with respective prioritized anatomical regions a difficult task and exceeds the scope of this research. The following are some examples for classifying medical form text into categories (see Table 2):

*Example 1:* A patient treated for an emergency related to her pregnancy would be included in the *Reproductive System* category (see Table 2).

*Example 2:* A conscious and breathing patient treated for gun shot wounds to the abdominal region would fall into the *Circulatory/Cardiovascular System* due to potential loss of blood, as well as being categorized for *Abdominal, Back, and Pelvic* categories (see Table 2).

We take characters with the highest recognition as an input and produce higher level topic categories. A knowledge base is constructed during the *training phase* from a set of PCR forms. The knowledge base contains the relationships between terms and categories. The *recognition phase* takes

---

<i>10 Body Systems:</i> Circulatory/Cardiovascular, Digestive, Endocrine, Excretory, Immune, Integumentary, Musculoskeletal, Nervous, Reproductive, Respiratory.
<i>6 Body Range Locations:</i> Abdomen, Back/Thoracic/Lumbar, Chest, Head, Neck/Cervical, Pelvic/Sacrum/Coccyx.
<i>4 Extremity Locations:</i> Arms/Shoulders/Elbows, Feet/Ankles/Toes, Hands/Wrists/Fingers, Legs/Knees.
<i>4 General:</i> Fluid/Chemical Imbalance, Full Body, Hospital Transfer/Transport, Senses.

---

**Table 2** Categories are denoted by these Anatomical Positions

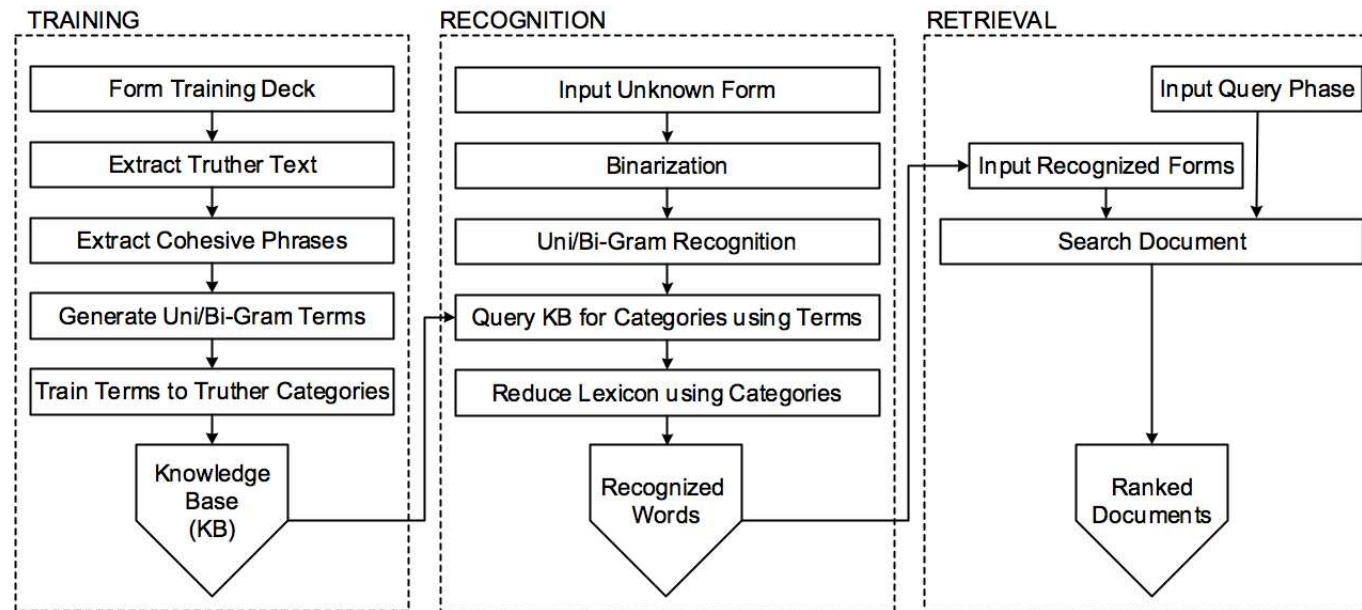
an unknown form, and reduces the lexicon using the knowledge base. This phase is evaluated using a separate testing set. Finally, after all content of the PCR form has been recognized, a search can take place by entering in a query. This phase is tested by querying the system with a set of phrase inputs. The forms are then ranked accordingly and returned to the user. The complete architecture of the proposed algorithm is also shown in the Figure 2.

In the training phase, a mechanism for relating uni-grams and bi-grams (henceforth: uni/bi-grams) as well as categories from a PCR training set are constructed. The testing phase then evaluates the algorithm’s ability to determine the categories from a test form by using a lexicon driven word recognizer (LDWR) [36] to extract the top-choice uni/bi-gram characters from all words. A maximum of two characters per word are considered, given that LDWR [36] successfully extracts a bi-gram with spatial encoding information 40% of the time. If  $\geq 3$  characters are selected, then LDWR [36] successfully extracts a character  $\leq 1\%$  of the time. Hence the maximum value of  $n$  in the  $n$ -grams is taken to be 2 (see examples in Figure 4).

### 3.1 Training

The training stage involves a series of steps to construct a matrix that represents relationships between terms and categories. Each form can have up to five categories. In the first phase, lexicons are constructed using all the words from all forms under a category. In the second phase, phrases are extracted from the form using a cohesion equation. These phrases are then converted to ESI encoding terms (ESI denotes “Exact Spatial Information” used as the encoding procedure for the uni/bi-gram terms; see definitions later in this section). A matrix is then constructed utilizing the ESI terms for the rows and the categories in the columns. The matrix is then normalized, weighted, and prepared in Singular Value Decomposition format.

A list of about 400 stopwords provided by PubMed are omitted from the lexicon [49] [29]. An additional list of about 50 words (e.g. male, female, etc.) found in most PCR’s, which have little bearing on the category are omitted from the cohesion analysis (the frequency of two words co-occurring



**Fig. 2** Proposed Algorithm Road Map

versus occurring independently; see Equation 1) but retained in the final lexicon. The term extraction procedure is also shown in the Figure 3. It is also common to apply other filters to reduce the likelihood of morphological mismatches [29]. Finally, word stemming is applied after the LDWR [36] has determined the ASCII word translation.

A passage  $P$  is the set of all words  $w$  for a PCR form under a category  $C$  treated as a single string. For each  $C$ , every pair of passages, denoted  $P_1$  and  $P_2$ , is compared. A phrase is defined as a sequence of adjacent non-stopwords [19]. Here we denote  $w_x$  as a word located at position  $x$  within a passage  $P$ . Let  $a, b$  and  $a', b'$  denote the index of words in an ordered passage  $P_1$  and  $P_2$  respectively ( $w_a \in P_1, w'_a \in P_2, w_b \in P_1, w'_b \in P_2$  such that  $b' > a'$  and  $b > a$ ) then a potential phrase consisting of exactly two words is constructed. The cohesion of phrases under each  $C$  is then computed. If the cohesion is above a threshold, then that phrase represents that category  $C$ . Thus a category  $C$  is represented by a sequence of high cohesion phrases using only those PCR passages manually categorized under  $C$ .

$$cohesion(w_a, w_b) = z \bullet \frac{f(w_a, w_b)}{\sqrt{f(w_a)f(w_b)}} \quad (1)$$

The cohesion between any two words  $w_a$  and  $w_b$  is computed by the frequency that  $w_a$  and  $w_b$  occur together versus existing independently. The top 40 cohesive phrases are retained for each category (see Equation 1). In the given equation,  $z$  is a constant ( $z = 2$  in the current research). The idea here is to analyze relationships between two words based on their correlations. If the two words are related to a concept in some way, a higher correlation measure would reflect it accordingly.

Consider the following two unfiltered strings of words  $S_1$  and  $S_2$  under the category *legs*:

$S_1$ : "right femur fracture"

$S_2$ : "broken right tibia and femur"

The candidate phrases  $CP_1$  and  $CP_2$  after the filtering step are:

$CP_1$ : "right femur" ... "right fracture" ... "femur fracture"

$CP_2$ : "broken right" ... "right femur" ...

The phrase "right femur" is computed from  $CP_1$  and  $CP_2$ , given that  $w_a$  and  $w'_a = \text{"right"}$ ,  $w_b$  and  $w'_b = \text{"femur"}$ , and the conditions  $b > a$  and  $b' > a'$  have been met. If the cohesion for "right femur" is above the threshold across all PCR forms under the *legs* category, then this phrase is retained as a representative of the category *legs*.

Tables 3 and 4 illustrate some top choice cohesive phrases generated. Digestive system and pelvic region are anatomically *close*. However, different information is reported in these two cases resulting in mostly different cohesive phrases. Those which are the same, such as *CHEST PAIN* have different cohesion values. This implies that it is likely that the term frequencies will also be different and therefore commonly occurring terms need to be weighted appropriately to their categories (this will be discussed in more detail later). Phrases sometimes may not make sense by themselves, however,

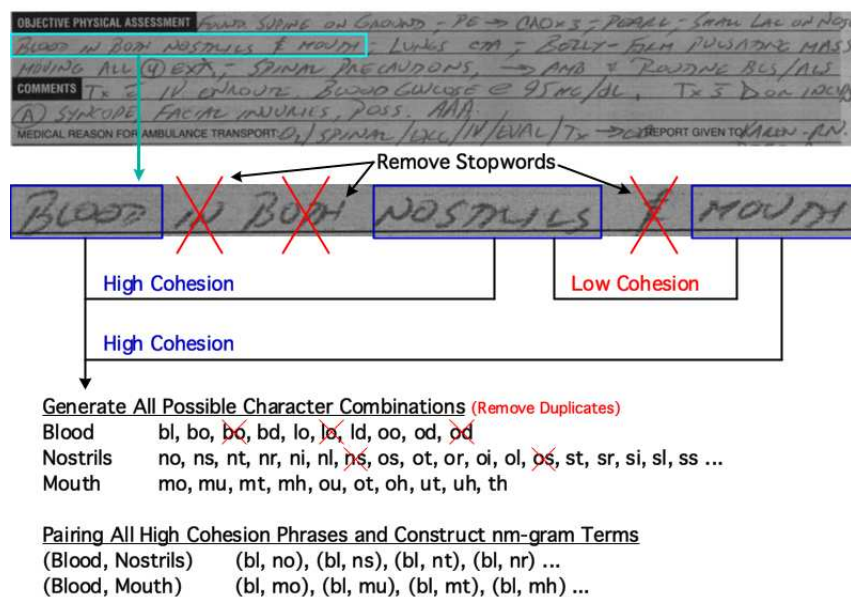


FREQUENCY	COHESION	PHRASE
6	0.67	DCAP BTLS
166	0.35	CHEST PAIN
91	0.38	PAIN 0
1860	2.49	PAIN HIP
144	0.34	HIP JVD
112	0.39	PAIN CHANGE
275	0.81	HIP FX
110	0.37	HIP CHANGE
82	0.38	PAIN 10
163	0.40	JVD PAIN
106	0.40	CAOX3 PAIN
202	0.50	PAIN JVD
213	0.55	PAIN LEG
205	0.42	CHEST HIP
3	0.33	PERPENDICULAR DECREASE
121	0.33	FELL HIP
118	0.36	PAIN FX
2251	3.01	HIP PAIN
390	0.83	PAIN CHEST
288	0.59	HIP CHEST

**Table 3** Top Cohesive Phrases for the Category: Pelvis

FREQUENCY	COHESION	PHRASE
30	0.72	PAIN INCIDENT
5	0.31	PAIN TRANSPORTED
42	0.54	PAIN CHEST
52	0.81	STOMACH PAIN
9	0.25	HOME PAIN
6	0.43	VOMITING ILLNESS
39	0.51	CHEST PAIN
4	0.24	CHEST SOFT
25	0.54	PAIN SBM
31	0.37	PAIN X4
31	0.47	PAIN JVD
11	0.34	PAIN EDEMA
25	0.44	PAIN PMSX4
6	0.21	PAIN SOFT
3	0.21	SBM INCIDENT
11	0.25	PAIN LEFT

**Table 4** Top Cohesive Phrases for the Category: *Digestive System*



**Fig. 3** Term Extraction from High Cohesive Phrases

this is the result of using a cohesive phrase formula in which words may not be adjacent.

There are three strategies for term representations: NSI, ESI and ASI. These terms will later be modeled to an anatomical category and used as the essential criterion for lexicon reduction.

### No Spatial Information (NSI):

An asterisk (\*) indicates that zero or more characters are found between  $C_1$  and  $C_2$ . NSI encodings are the most simple form of encoding (see Figure 4 examples).

UNI-GRAM ENCODING:  $*C*$

BI-GRAM ENCODING:  $*C_1 * C_2*$

BI-GRAM ENCODING EXAMPLE: BLOOD →  $*L*D*$

### Exact Spatial Information (ESI):

The integers (x, y, z) represent the precise number of characters between  $C_1$  and  $C_2$ . ESI encodings are an extension of the NSI encodings with the inclusion of precise spatial information. In other words, the number of characters before, after and between the highest confidence  $C_1$  and  $C_2$  characters are part of the encoding. These encodings are the most successful in our experiments since there are fewer term collisions involved. Hence the ESI encodings are preferred.

UNI-GRAM ENCODING:  $x C y$

BI-GRAM ENCODING:  $x C_1 y C_2 z$

(ID: 342-2)		*C*S* *A*N*
(ID: 407-1)		*C*M* *P*A*
(ID: 407-2)		*H*O* *A*W*
(ID: 473)		*C*F* *L*A*
(ID: 643)		*C*H* *A*W*
(ID: 695-1)		*L*S* *L*R*
(ID: 98)		*C*S* *S*S*
(ID: 606)		*C*W* *D*M*
(ID: 695-2)		*C*F* *W*L*

**Fig. 4** NSI Encodings Example (Blue Letters: LDWR[36] successfully extracted)

BI-GRAM ENCODING EXAMPLE: BLOOD  $\rightarrow$  1L2D0

#### Approximate Spatial Information (ASI):

The integers  $(x_a, y_a, z_a)$ , denoted as length codes, represent an estimated range of characters between  $C_1$  and  $C_2$ . A '0' indicates no characters, a '1' indicates between one and two characters, and a '2' represents greater than 2 characters. The ASI encodings are an approximation of ESI encodings designed to handle cases when the precise number of characters is not known with high confidence.

UNI-GRAM ENCODING:  $x_a C y_a$

BI-GRAM ENCODING:  $x_a C_1 y_a C_2 z_a$

BI-GRAM ENCODING EXAMPLE: BLOOD  $\rightarrow$  1L1D0

#### Combinatorial Analysis

The quantity of all possible NSI, ESI and ASI uni-gram and bi-gram combinations, for a given word of character length  $n$ , such that  $n \geq 1$ , is represented by Equation 2. Regardless of the encoding, the same quantity of combinations exists since the distance between characters is known.

$$\mathcal{C}(n) = \left( \left( \sum_{i=1}^{n-1} (n-i) \right) + n \right) = \left( \left( \left( \frac{n}{2} \right) (n-1) \right) + n \right) \quad (2)$$

However, the function  $\mathcal{C}$  only considers the combinations of an individual entry. The combination inflation of a uni/bi-gram phrase is shown by

Equation 3. The equation parameters  $a$  and  $b$  represent the string lengths of the words considered in a phrase. The total number of possible uni/bi-gram combinations resulting from a phrase  $P$  containing two words of length  $a$  and  $b$  is the product of the possible combinations of each word denoted as  $\mathcal{C}(a)$  and  $\mathcal{C}(b)$  respectively.

$$\mathcal{P}(a, b) = \mathcal{C}(a) \cdot \mathcal{C}(b) \quad (3)$$

For example:

Let the phrase to evaluate uni/bi-gram combinations be *PULMONARY DISEASE*.

Let  $n = \text{length}(\text{"PULMONARY"}) = 9$

Let  $m = \text{length}(\text{"DISEASE"}) = 7$

$\mathcal{C}(n) = 45$  uni-gram + bi-gram combinations for "PULMONARY"

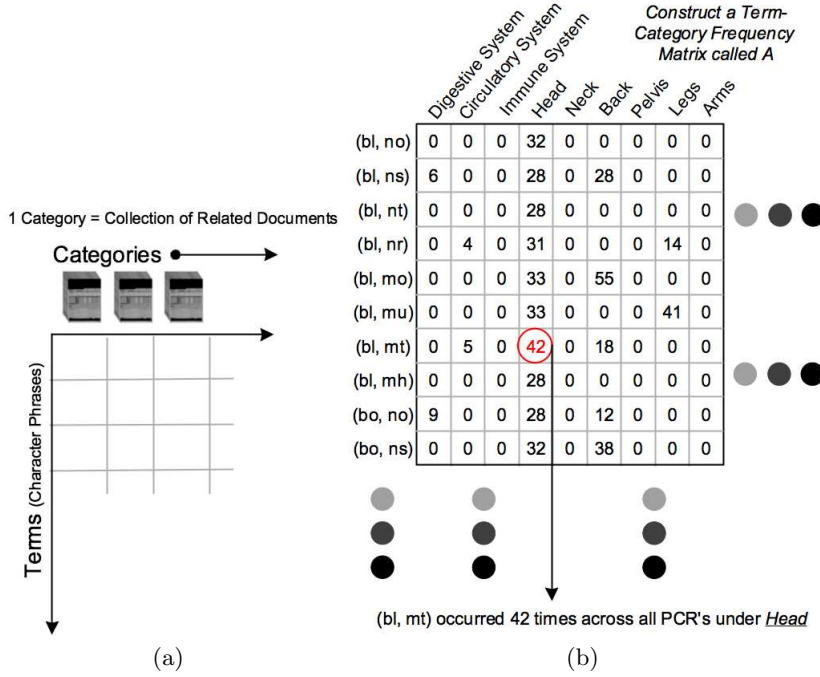
$\mathcal{C}(m) = 28$  uni-gram + bi-gram combinations for "DISEASE"

$\mathcal{P}(n, m) = 1,260$  uni-gram + bi-gram phrase combinations for *PULMONARY DISEASE*

Each of these encodings has its advantages and disadvantages. The choice is ultimately based on the quality of the handwriting recognizer's (LDWR) ability to extract characters. If the handwriting recognizer cannot successfully extract positional information, then NSI is the best approach. If extraction of positional information is reliable, then the ESI is the best approach. However, NSI and ASI create more possibilities for confusion since distances are either approximated or omitted. ESI is more restrictive on the possibilities as the precise spacing is used leading to lesser confusion among terms.

Using the ESI protocol, all possible uni/bi-gram terms are synthetically extracted from each cohesive phrase under each category. For example, BLOOD can be encoded to the uni-gram 0B4 (zero characters before 'B' and four characters after 'B') and the bi-gram 0B3D0 (zero characters before 'B', three characters between 'B' and 'D' and zero characters following 'D'). All possible synthetic positional encodings are generated for each phrase and chained together (a '\$' is used to denote a chained phrase). For example, CHEST PAIN encodes to: 0C4\$0P0A2 ... 0C4\$1A2 ... 0C0H3\$0P1I1 ... 0C0H3\$0P2N0, etc. Therefore, each category now has a list of encoded phrases consisting of positional encoded uni/bi-grams. These terms are the most primitive representative links to the category used throughout the training process. In the training phase, the synthetic information can be extracted since the text is known. However, in the testing phase, a recognizer will be used to automatically produce an ESI encoding since the test text is not known. To improve readability, the notation  $(W_1, W_2)$  is used to represent an ESI encoding of a two-word phrase (e.g. Myocardial Infarction: (my, in), (my, if), (my, ia), etc ...).

A matrix  $A$ , of size  $|T|$  by  $|C|$ , is constructed such that the rows of the matrix represent the set of terms  $T$ , and the columns of the matrix represent the set of category  $C$  as shown in Figure 5(a). The value at matrix coordinate  $(t, c)$  is the frequency that each term is associated with the category. The term frequency corresponds to the phrasal frequency from which it was



**Fig. 5** (a) Term Category Matrix (TCM) Overview (b) TCM frequency construction Example

derived. It is the same value as the numerator in the cohesion formula (refer to Equation 1):  $f(w_a, w_b)$ . For example, if the frequency of CHEST PAIN is 50, then all term encodings generated from CHEST PAIN, such as (ch, pa), will also receive a frequency of 50 in the matrix. An example of term frequency construction is shown in Figure 5(b).

**Step 1:** Compute the normalized matrix B from A using Equation 4 [9] [10], where normalisation for a term is done over all possible categories:

$$B_{t,c} = \frac{A_{t,c}}{\sqrt{\sum_{e=1}^n A_{t,e}^2}} \quad (4)$$

Matrix A is the input matrix containing raw frequencies, Matrix B is the output matrix with normalized frequencies, and (t,c) is a (term, category) coordinate within a matrix. The normalisation equation is used to normalise the frequency count of a term in a given category by the frequency of the same term in all possible categories, which reflects how representative the term is with respect to the given category.

**Step 2:** Term Discrimination Ability

The Term Frequency times Inverse Document Frequency (TF x IDF) is used to favor those terms which occur frequently with a small number of categories as opposed to their existence in all categories [41] [59]. While Luhn [41] asserts that medium frequency terms would best resolve a document, it precludes classification of rare medical words. Salton's [59] theory, stating

that terms with the most discriminatory power are associated with fewer documents, allows a rare word to resolve the document.

STEP 2A Compute the weighted matrix X from B using Equation 5 [9] [10] [29]:

$$IDF(t) = \log_2 \frac{n}{c(t)} \quad (5)$$

IDF gives the inverse-document-frequency on term  $t$ , where  $c(t)$  is the number of categories containing term  $t$ .

Step 2B Weight the normalized matrix by IDF values using Equation 6 [9] [10] [32] [29]:

$$X_{t,c} = IDF(t) \cdot B_{t,c} \quad (6)$$

Matrix B is the normalized matrix from Step 1, IDF is the computational step defined in Step 2, and Matrix X is a normalized and weighted matrix.

The normalized and weighted term-category matrix can now be used as the knowledge base for subsequent classification. A singular value decomposition variant, which incorporates a dimensionality reduction step allows a large term-category matrix to represent the PCR training set (see Equation 7). This facilitates a category query from an unknown PCR using the LDWR [36] determined terms [9] [10] [15].

$$X = U \bullet S \bullet V^T \quad (7)$$

Matrix X is decomposed into 3 matrices: U is a (T x k) matrix representing term vectors, S is a (k x k) matrix, and V is a (k x C) matrix representing the category vectors.

The value k represents the number of dimensions to be finally retained. If k equals the targeted number of categories to model, then SVD is performed without the reduction step. Therefore, in order to reduce the dimensionality, the condition  $k < |C|$  is necessary to reduce noise [15].

### 3.2 Testing

Given an unknown PCR form, the task is to determine the category of the form, and use the reduced lexicon associated with the determined category to drive the word recognizer, LDWR [36]. In addition, the category determined can be used to tag the form which can be subsequently used for information retrieval. The query task is divided into the following steps: (i) Term Extraction, (ii) Pseudo-Category Generation, and (iii) Candidate Category Selection [9] [10].

Given a new PCR image, all image words are extracted from the form, and LDWR [36] is used to produce a list of confidently recognized characters for each word. These are used to encode the positional uni/bi-grams consistent with the format during training. All combinations of uni/bi-phrases in the PCR form are constructed. Each word has exactly one uni-gram and

exactly one bi-gram. A phrase consists of exactly two unknown words. Therefore it is represented by precisely four uni/bi-phrases (BI-BI, BI-UNI, UNI-BI and UNI-UNI).

A  $(m \times 1)$  query vector  $Q$  is derived, which is then populated with the term frequencies for the generated sequences from the Term-Extraction step. If a term is not encountered in the training set, then it is not considered. Positional bi-grams are generated to yield the trained terms 37% of the time, and similarly positional uni-grams 57% of the time. The experiments here illustrate this to be a sufficient number of terms. A scaled vector representation of  $Q$  is then produced by multiplying  $Q^T$  and  $U$ .

Once the pseudo category is derived, R-SVD is applied for the following reasons: (i) It converts the query into a vector space compatible input, and (ii) the dimensional reduction can help reduce noise [15]. Since the relationship between terms and categories is scaled by variance, the reduction allows parametric removal of less significant term-category relationships.

The task is now to compare the pseudo-category vector  $Q$  with each vector in  $V_r \bullet S_r$  (from the training phase) using a scoring mechanism. The cosine rule is used for matching the query [9] [10]. Both  $x$  and  $y$  are dimensional vectors used to compute the cosine in Equation 8. Vectors  $x$  and  $y$  in the equations represent the comparison of the vectors: pseudo-category  $Q$  with every column vector in  $V_r \bullet S_r$ .

$$z = \cos(x, y) = \frac{x \cdot y^T}{\sqrt{\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i^2}} \quad (8)$$

Each cosine score is mapped onto a sigmoid function using the least square fitting method, thereby producing a more accurate confidence score [9] [10]. The least squares regression line used to satisfy the equation  $f(x) = ax + b$  are shown in Equations 9 and 10 [39]:

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad (9)$$

$$b = \frac{1}{n} \left( \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i \right) \quad (10)$$

The fitted sigmoid confidence is produced using the cosine score and the regression line, using equation (9):

$$\text{confidence}(a, b, z) = \frac{1}{1 + e^{-(az+b)}} \quad (11)$$

The confidence scores are then used to rank the categories. If a category is above an empirically chosen threshold, then that category is retained for the PCR. Multiple categories may be thus retained. All words corresponding to the selected categories are then used to construct a new lexicon which is finally submitted to the LDWR recognizer [36]. Given a test PCR form, and the reduced lexicon, the LDWR [36] converts the handwritten medical words to ASCII.

Each word which is recognized is compared with the truth. However, a simple string comparison is insufficient due to spelling mistakes and root variations of word forms which are semantically identical. This occurs 20% of the time within the test set words. Therefore, Porter stemming [52] [33] [57] and Levenshtein String Edit Distance [4] of 1 allowable penalty are performed on both the truth and the recognizer result before they are compared. Levenshtein is only applied to a word that is believed to be  $\geq 4$  characters in length. For example, PAIN and PAINS are identical. However, this also results in an improper comparison in about 11% of the corrections.

## 4 Recognition Experiments

Our training data consists of 750 PCR forms and the test data consists of a separate blind set of 62 PCR forms. In all experiments it is assumed that the word segmentation and extraction has been performed by a person. Also, forms in which 50% of the content is indecipherable by a human being are omitted. This occurs 15% of the time. A description of the training and test sets can be found in Table 5.

ENVIRONMENT ITEM	VALUE
Training Set PCR Size	750
Testing Set PCR Size	62
Training Set Lexicon Size	5,628
Testing Set Lexicon Size	2,528
Training + Testing Set Lexicon Size	8,156
Training Set Words for Modeling	42,226
Testing Set Words to Recognize	3,089
Modeled Categories / RSVD Dimensions	24
Category Selection Threshold	0.55
Maximum Categories per Form	5
Average Categories per form	1.40
Max Phrases Per Category	50
Apple OS X Memory Usage	520 MB
Apple OS X G4 1GHZ Train Time	15-20 mins/exp
Apple OS X G4 1GHZ Test Time	3 hrs/exp

**Table 5** Handwriting Recognition System Environment

### 4.1 Performance Measures

Table 6 contains 6 rows corresponding to performance measure in recognition performance. These fields are ACCEPT, ERROR, RAW, LEXICON-SIZE, NOT-IN-LEXICON and HARD-TO-READ which are explained as follows:



*ACCEPT (accept recognition rate)*: number of words the word recognizer accepts above an empirically decided threshold.

*ERROR (error recognition rate)*: number of words incorrectly recognized among the accepted words.

*RAW (raw recognition rate)*: top choice word recognition rate without use of thresholds.

*LEXICON-SIZE (lexicon size)*: the lexicon size for the experiment after any reductions.

*NOT-IN-LEXICON (truth word not present in the lexicon)*: percentage of words (for a specific experiment) not in the lexicon as a result of incorrectly chosen categories or due to the absence of that word in the training set.

*HARD-TO-READ (human being could not completely decipher word)*: percentage of the NOT-IN-LEXICON set in which even human beings could not reliably decipher all or some of the characters in the word (given the context).

Table 7 contains conclusions in raw recognition and error rate based on the experiments in Table 6. These fields are RAW-RATE and ERROR-RATE which are explained as follows:

*RAW-RATE*: shows the improvement (denoted by an upward arrow in Table 7) in raw recognition performance between experiments.

*ERROR-RATE*: shows the reduction (denoted by a downward arrow in Table 7) in the incorrect accept rate between experiments.

## 4.2 Experiments

This section describes several kinds of experiments which correspond to Table 6. The purpose of these experiments is to compare and contrast the theoretical maximum recognition performance with the actual recognition performance. There are 4 major types of experiments: (C)omplete, (A)ssumed, (R)educed, and (S)ynthetic. The complete experiment means the recognizer was executed with the full lexicon. The assumed experiment means that a theoretically reduced lexicon is constructed under the assumption that the medical form categories are supplied by an oracle. The reduced experiment means that the actual latent semantic analysis in this paper is used to extract a reduced lexicon from recognized medical form categories. The synthetic experiment means that the uni/bi-grams were theoretically known (i.e. the handwriting recognizer always extracted 2 characters with 100% accuracy). However, since all words in a test set may not have been seen in a training set, the 4 experiments are executed in two modes: (i) with just words from the training set, and (ii) words merged from both the training and testing sets. These two modes allow us to compare the performance in situations of known versus unseen words in a form. To indicate in the charts the different of each of 4 experiments in 2 modes, we use acronyms: CL and CLT for complete lexicon analysis in mode 1 and 2 respectively, and similarly AL vs. ALT, SL vs. SLT, and finally RL vs. RLT. The experimental results can be found in Tables 6 and 7 with discussion that follows.

	CL	CLT	AL	ALT	SL	SLT	RL	RLT
<i>ACCEPT</i>	76.34%	76.92%	63.52%	66.59%	70.51%	71.51%	70.70%	71.06%
<i>ERROR</i>	71.93%	69.65%	57.24%	47.12%	62.26%	59.44%	62.04%	59.45%
<i>RAW</i>	23.31%	25.32%	32.31%	41.73%	30.30%	32.73%	30.62%	32.63%
<i>LEXICON-SIZE</i>	5,628	8,156	1,193	1,246	2,514	2,620	2,401	2,463
<i>NOT-IN-LEXICON</i>	-	-	23.89%	8.02%	16.06%	10.46%	16.61%	12.23%
<i>HARD-TO-READ</i>	-	-	33.33%	97.98%	48.19%	73.99%	46.59%	62.96%

**Table 6** Handwriting Recognition Performance

	CLT to RLT	CL to RL	CLT to ALT	CLT to SLT
RAW Rate	↑ 7.48%	↑ 7.42%	↑ 17.58%	↑ 7.42%
Error Rate	↓ 10.78%	↓ 10.88%	↓ 24.53%	↓ 10.21%

**Table 7** Comparison between Handwriting Recognition Experiments

### 4.3 Discussion

In reference to Table 7 which is computed from the most relevant changes of Table 6: The theoretical RLT (i.e. comparing RLT to CLT) improves the RAW match rate by 7.48% and drops the error rate 10.78% with a *degree of reduction*  $\rho = 61.59\%$ . The practical RL (i.e. comparing RL to CL) improves the RAW match rate by 7.42% and drops the error rate by 10.88%. The RLT and RL numbers are close due to the difference in the initial lexicon sizes: CLT/RLT starts with 6,561 words (i.e. training set and testing set lexicons) whereas the CL/RL starts with 5,029 words (i.e. training set lexicon only). The RLT lexicon is more complete, but the lexicon is larger. The RL lexicon is less complete, but the lexicon is smaller. Thus, RLT gives the advantage that the recognizer has a greater chance of the word being a possible selection and RL gives the advantage of the lexicon being smaller. The ALT shows the theoretical upper bound for the paradigm: (i) the categories are correctly determined 100%, and (ii) the lexicon is complete. The ALT (i.e. comparing ALT to CLT) improves the RAW match rate by 17.58% and drops the error rate 24.53% with a *degree of reduction*  $\rho = 83.01\%$ . The synthetic experiments (SL and SLT) also do not offer much improvement which shows perfect character extraction does not guarantee recognition improvement. This is due to two reasons: (i) a form is a representation of many characters and so some incorrectly recognized characters are tolerated, and (ii) the remaining words on the form to be recognized are difficult to determine even when the lexicon is constructed with only the words of known uni/bi-gram terms.

Table 8 provides insight into the effectiveness of the lexicon reduction from the complete lexicon (CL) to the reduced lexicon (RL) experiments. The performance measures for lexicon reduction as described by Madhvanath [42] and Govindaraju, et al. [26] are used with alteration to the definition of reduc-

LEXICON ANALYSIS METRIC	VALUE
Accuracy of Reduction ( $\alpha$ )	0.33
Degree of Reduction ( $\rho$ )	0.83
Reduction Efficacy ( $\eta$ )	0.06
Lexicon Density ( $\varrho'$ )	1.07 $\rightarrow$ 0.87
Lexicon Density ( $\varrho''$ )	0.50 $\rightarrow$ 0.78

**Table 8** Lexicon Reduction Performance between the Complete Lexicon (CL) and the Reduced Lexicon (RL)

tion efficacy. The *Accuracy of Reduction*  $\alpha = E(\mathcal{A})$ , such that  $\alpha \in [0, 1]$  [42], and  $\mathcal{A}$  is a random variable [5], indicates the existence of the truth in the lexicon. The function  $E$  computes the expectation [5]. The *Degree of Reduction*  $\rho = E(\mathcal{R})$ , such that  $\rho \in [0, 1]$  [42], represents the mean size of the reduced lexicon. The *Reduction Efficacy*  $\eta = \Delta_{LDWR} \times \alpha^{1-\rho}$ , such that  $\Delta_{LDWR}$ ,  $\eta, \alpha, \rho \in [0, 1]$ , is a measure of the effectiveness of a lexicon with respect to a lexicon driven recognizer. This formula is defined differently in this research to weigh the importance of accuracy over the reduction and include the reductions effect on the recognizer. The larger the efficacy value is, the better is the effectiveness of the reduction for one recognizer versus another. The larger the *Lexicon Density*  $\varrho_{LDWR}(\mathcal{L}) = (v_{LDWR}(\mathcal{L}))(f_{LDWR}(n) + \delta_{LDWR})$  value (such that  $v_{LDWR}(\mathcal{L}) = \frac{n(n-1)}{\sum_{i \neq j} d_{LDWR}(\omega_i, \omega_j)}$  and  $d_{LDWR}(\omega_i, \omega_j)$  is a recognizer dependent computation used to denote a distance metric between two supplied words) the more *similar* or *close* the lexicon words are [26]. A supplemental distance measure denoted by the *N-Gram Lexicon Distance Metric*  $d_{LDWR}(\omega_i, \omega_j) = \gamma(\omega_i, \omega_j) / \Gamma(\omega_i, \omega_j)$ , introduced in this research and substituted into the lexicon density equation  $\varrho$ , provides a measure of uni/bi-grams existing within the lexicon. The value  $\gamma$  represents the number of uni/bi-gram terms that are *not* common between  $\omega_i$  and  $\omega_j$ .  $\Gamma$  denotes the total number of uni/bi-gram term combinations between  $\omega_i$  and  $\omega_j$ . In order to distinguish between the *lexicon density distance metric* and the *n-gram lexicon distance metric* equations, the values  $\varrho'$  and  $\varrho''$  will be respectively used. The *lexicon density distance metric*  $\varrho'$  shows less confusion among lexicon words considering all the characters are equally important. This implies that the reduced lexicon will be less confusing to the recognizer. The *n-gram lexicon distance metric* shows an increase in the quantity of words with common NSI encodings. This implies the recognizer has a greater chance of selecting a word using the confidently selected characters.

## 5 Search Experiments

The ability to query a set of PCR medical forms which match a user supplied input phrase is important for Health Surveillance applications. Searching text in digital format is easily accomplished but this is much harder to do for scanned handwritten documents. While searching handwriting has only been demonstrated in certain areas [56]. The experiments in this section illus-

trate search effectiveness even when words are incorrectly recognized. Both the original LDWR (CL) and the reduced lexicon LDWR (RL) PCR medical form data sets are compared.

In order to have a query set of sufficient size, the test set is constructed using a leave-1-out strategy. There are 8 rounds of recognition such that each round of the 800 PCR's are divided into two different groups of 100 and 700. During each of the rounds, the content of the 100 PCR's is recognized using the 700 PCR's as the training data. This allows the full set to be evaluated with no bias. Finally, a set of 1175 phrases, constructed from adjacent non-stopwords, are extracted from a blind set of 200 PCR forms (i.e. these 200 forms are not a subset of the 800 set) such that each phrase is found in at least one form in the 800 set. Each of the query phrase in the query set consists of exactly two words. Different experiments are conducted which search the PCR forms for at least one of the words or both the words from the input query phrase.

A query is performed by scanning the forms in the 800 test set for recognized words that match a two-word input query phrase. Any LDWR recognized form which contains the occurrence of both query words independently in the document are considered matched results. Relevancy is determined if the input query words, for example *CHEST* and *PAIN*, are actually found on that form according to the human truth. A two-step ranking algorithm is then performed on all matching documents. First documents are ranked according to the frequencies of the occurring words. Second, those documents with the same word frequency are ranked using the distance measurement in Equation 12. Let  $d(a_i, b_j)$  be a function which computes the distance between two matched words,  $a_i$  and  $b_j$  such that  $i$  and  $j$  respectively represent the word position in the document.  $w_{ij}$  here is a weight based on the frequency of occurrences of words  $a$  and  $b$  in the document. This is especially necessary in situations where word  $a$  exists and  $b$  does not, and vice versa. Documents with closer proximity words are given a higher rank. Discussion on proximity based metrics can be found here [23]. Finally, the search methods are evaluated using the standard trec\_eval system. To account for cases, where the system improperly returns no documents for a given query, -c option of trec\_eval is used to include the relevance count of these queries in the final calculation.

$$d(a_i, b_j) = w_{ij} * \frac{1}{|a_i - b_j|} \quad (12)$$

### 5.1 Performance Measures

*MAP (mean average precision)*: It is the mean of the average precision of all individual queries in the set. Average precision of a single query is defined as the mean of the precision after every relevant document retrieved. This performance measure emphasises on retrieving relevant documents earlier.

*R-prec (R-precision)*: R-precision is the precision at  $R$ , where  $R$  denotes total

number of relevant documents for the given query. This measure emphasises on retrieving more relevant documents.

## 5.2 Experiments

*AND CL* : Given a query phrase of two words, both words are found in a PCR form during the search process using a complete training lexicon.

*AND RL* : Given a query phrase of two words, both words are found in a PCR form during the search process using a reduced training lexicon.

*OR CL* : Given a query phrase of two words, at least one of the words is found in a PCR form during the search process using a complete training lexicon.

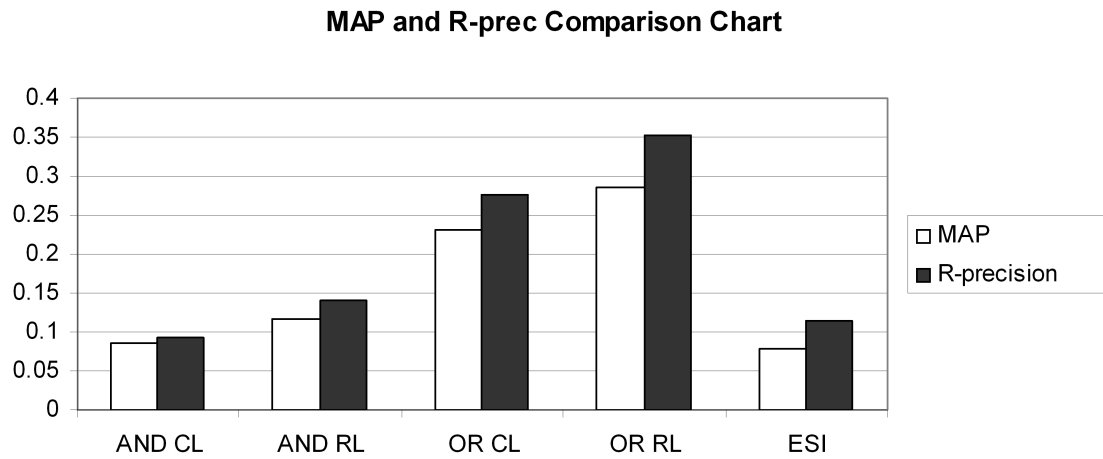
*OR RL* : Given a query phrase of two words, at least one of the words is found in a PCR form during the search process using a reduced training lexicon.

*ESI* : An additional query expansion experiment was also performed in which a document was matched if at least one ESI encoding sequence was found in the document (i.e. the requirement for matching words was removed). For example, consider input query phrase *CHEST PAIN* where *CHEST* is decomposed into CH, CE, CS, CT, HE, HS, HT, ES, ET, C, H, E, S, and T., and *PAIN* is decomposed into PA, PI, PN, AI, AN, IN, P, A, I, and N. Since the input phrase is known, and hence the spatial encodings between characters are also known, the ESI encodings for the terms are known. The ESI encodings for *CHEST* are decomposed into: 0C0H3, 0C1E2, 0C2S1, 0C3T0, 1HE2, 1H1S1, 1H2T0, 2E0S1, 2E1T0, 0C4, 1H3, 2E2, 3S1, and 4T0. The ESI encodings for *PAIN* are decomposed into: 0P0A2, 0P1I1, 0P2N0, 1A0I1, 1A1N0, 2I0N0, 0P3, 1A2, 2I1, and 3N0. Finally, all possible ESI sequences from the input words are generated: 0C0H3\$0P0A2, 0C0H3\$0P1I1, 0C0H3\$0P2N0, 0C0H3\$1A0I1, etc.

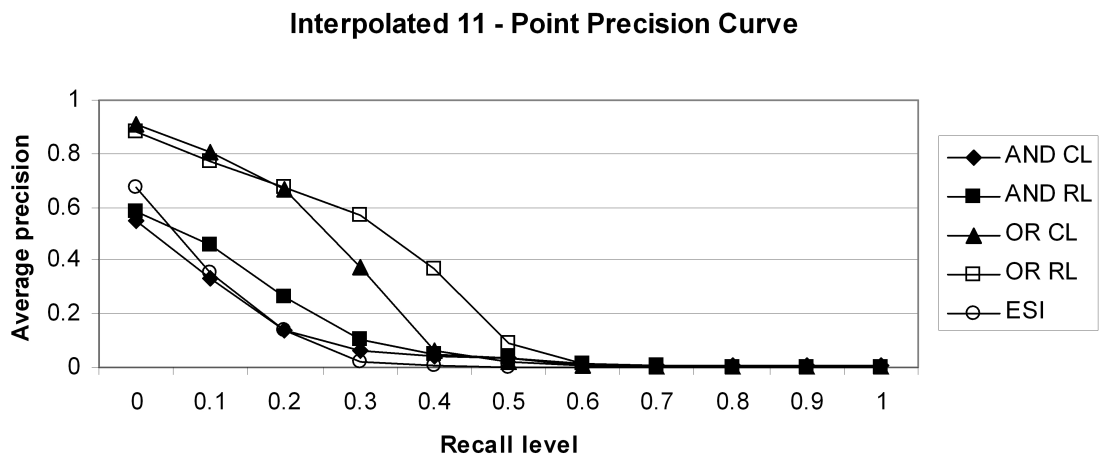
## 5.3 Discussion

The experimental results for each algorithm in terms of MAP and R-precision is shown in the Figure 6. As shown, retrieval based on reduced lexicon (RL) outperform retrieval based on complete lexicon (CL). This behavior is observed irrespective of the fact whether search is performed using both words of the query phrase (AND) or at least one of the words of the query phrase (OR). An interpolated 11 - point precision curve shown in Figure 7 also supports this observation. As shown in the figure, after a recall level of 0.2, OR-RL method retrieves relevant documents earlier in the order as compared to OR-CL method. In the case of AND logic, RL based method performs better than CL based methods at all recall levels. The improvement in the search performance due to lexicon reduction algorithm used highlights the effectiveness of the proposed method. For the query expansion experiment (ESI) as intuitively expected, the uni/bi-grams match more terms in the test set due to the loss in word information. The precision chart in Figure 7 illustrates this drop in retrieval effectiveness and shows that searches are

more effective at the word level rather than raw encoding level. Similar drop in performance is observed for the query expansion technique in Figure 6 and Figure 8



**Fig. 6** Mean Average Precision and R-Precision comparison for different algorithms



**Fig. 7** Interpolated 11 - point precision curve

To study the effect of different methods on the total number of relevant documents retrieved, we also compute the value of recall level for top  $k$  documents retrieved as shown in the Figure 8. The results from the Figure 8 suggest that reduced lexicon (RL) based methods not only retrieve relevant documents earlier, but also retrieve more relevant documents overall as compared to their counterpart complete lexicon (CL) based methods. The contribution of this research is that the lexicon reduction strategy (i.e. the RL experiment) improves both handwriting recognition and search effectiveness.

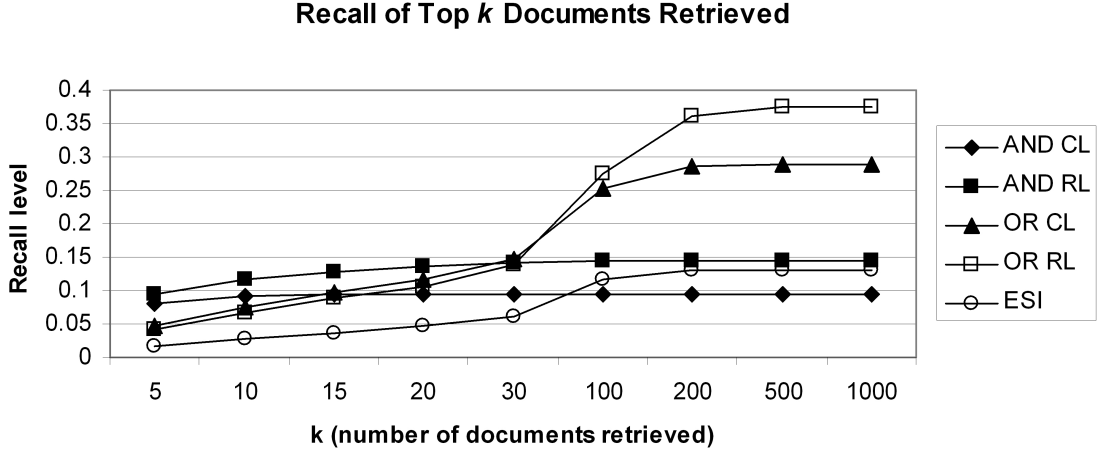


Fig. 8 Recall level of top  $k$  documents retrieved

## 6 Conclusions

This paper defines a new paradigm for lexicon reduction and information retrieval in the complex situation of handwriting recognition of medical forms. An improvement in raw recognition rate from about 25% of the words on a PCR form to approximately about 33% has been shown with a reduction in false accepts by about 7%, a reduction in error rate by about 10%-25%, and a lexicon reduction from 32%-85%. The addition of a category driven query facilitates a mean average precision of 0.28 and R-prec of 0.35 for 1175 queries in a search engine experiment with medical forms. Additionally, using a reduced lexicon for searching medical form also enables retrieving more relevant number of documents overall, as compared to complete lexicon search.

Interestingly, certain computational elements of bootstrapping, described in our work, are consistent with the human interpretation of ambiguous handwriting using contextual cues. Our methodology accomplishes this by modeling character terms as a higher level semantic concept which

mimics the human ability to recognize a word within context, when some characters are unknown.

## Acknowledgment

Special acknowledgements to (i) the National Science Foundation (NSF) for providing funding for this project, to (ii) Casey Czamara, of the Western Regional Emergency Medical Services (WREMS) program operating under the New York State Department of Health for providing necessary resources [67].

## References

1. Bahlmann, C., Haasdonk, B., Burkhardt, H. On-Line Handwriting Recognition with Support Vector Machines - A Kernel Approach. International Workshop on Frontiers in Handwriting Recognition. 2002.
2. Balasubramanian, A., Meshesha, M., and Jawahar, C.V. "Retrieval from Document Image Collections," in Proceedings of Seventh IAPR Workshop on Document Analysis Systems, 2006, pp. 1-12.
3. Bayer, T., Kressel, U., Mogg-Schneider, H., and Renz, I. Categorizing paper documents. *Computer Vision and Image Understanding*, 70(3):299-306, 1998.
4. Black, P.E., ed. "Levenshtein Distance". *Algorithms and Theory of Computation Handbook*; CRC Press LLC, from Dictionary of Algorithms and Data Structures, NIST, 1999.
5. Blum, J.R., Rosenblatt, J.I. Probability and Statistics. Chapter 4: Random Variables and Their Distributions; Chapter 6: Expectations, Moment Generating Functions, and Quantiles. W.B. Saunders Company. USA, 1972.
6. Blumenstein, M., Verma, S. A Neural Based Segmentation and Recognition Technique for Handwritten Words. *IEEE International Conference on Neural Networks*. 1998.
7. Byun, H., Lee, S.W. Applications of Support Vector Machines for Pattern Recognition: A Survey. *Lecture Notes in Computer Science*; Springer. 2002.
8. Caesar, T., Gloger, J.M., Mandler, E. Using Lexical Knowledge for the Recognition of Poorly Written Words. *Third International Conference on Document Analysis and Recognition*. Volume 2. pp. 915-918. 1995.
9. Chu-Carroll, J., and Carpenter, B. Dialogue Management in Vector-Based Call Routing. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*. pp. 256-262, 1999.
10. Chu-Carroll, J., and Carpenter, B., Vector-Based Natural Language Call Routing. *Computational Linguistics*. Vol. 25, No. 3, pp. 361-388, 1999.
11. Chen, M.Y., Jundu, A., Zhou, J. Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1994.
12. Cho, S.B., Kim, J.H. Applications of Neural Networks to Character Recognition. *Pattern Recognition*. 1991.
13. Cho, S.B. Neural-Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals. *IEEE Transactions on Neural Networks*. 1997.
14. Croft, B., Harding, S.M., Taghva, K., and Borsack, J. An evaluation of information retrieval accuracy with simulated OCR output. In *Proceedings of Symposium on Document Analysis and Information Retrieval*, pages 115-126, 1994.
15. Deerwester, S., Dumais, S.T., Furnas, G.Q., Landauer, and, T.K., Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*. 41(6):391-407, 1990.



16. Doermann, D. The indexing and retrieval of document images: a survey. *Computer Vision and Image Understanding*, 70(3):287-298, 1998.
17. Edwards, J., and Forsyth, D. "Searching for character models," in *Proc. of the 19th Annual Conf. on Neural Information Processing Systems*, Vancouver, Canada, 2005, pp. 331-338.
18. Edwards, J., Teh, Y. W., Forsyth, D., Bock R., Maire M., and Vesom, G. "Making Latin manuscripts searchable using (gHMM)'s," in *Proc. of the 18th Annual Conf. on Neural Information Processing Systems*, 2004 , pp. 385-392.
19. Fagan, J. The Effectiveness of a Non-Syntactic Approach to Automatic Phrase Indexing for Document Retrieval. *Journal of the American Society for Information Science*, 40: 115-132. 1989.
20. Favata, J.T. Offline General Handwritten Word Recognition Using an Approximate BEAM Matching Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*; 23 (9); pp1009-1021. 2001.
21. Feng, S.L. and Manmatha, R. Classification Models for Historic Manuscript Recognition. *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR)* 2005.
22. Gader, P.D., Keller, J.M., Krishnapuram, R., Chiang, J.H. *Neural and Fuzzy Methods in Handwriting Recognition*. 1997.
23. Goldman, R. Shivakumar, N., Venkatasubramanian, S., Garcia-Molina, H. Proximity Search in Databases. *IEEE Proceedings of the International Conference on Very Large Databases*, p.26-37. 1998.
24. Golub, G.B., Van Loan, C.E. *Matrix Computations*. 2nd Edition. John Hopkins University Press, 1989.
25. Govindaraju, V., Xue, H. Fast Handwriting Recognition for Indexing Historical Documents. *First International Workshop on Document Image Analysis for Libraries (DIAL)*. 2004.
26. Govindaraju, V., Slavik, P., and Xue, H. Use of Lexicon Density in Evaluating Word Recognizers. *IEEE Trans PAMI*, Vol. 24, No.6, p.789-800. 2002.
27. Guillevic, D., Nishiwaki, D., and Yamada, K. Word Lexicon Reduction by Character Spotting. *Seventh International Workshop on Frontiers in Handwriting Recognition*. Amsterdam. 2000.
28. Harding, S.M., Croft, W. B., and Weir, C. "Probabilistic retrieval of OCR Degraded textt using n-grams," in *Research and Advanced Technology for Digital Libraries*, 1997, 345-359.
29. Hersh, W.R. *Information Retrieval: A Health and Biomedical Perspective*. 2nd Edition. Springer-Verlag, New York, Inc. USA. 2003.
30. Howe, N.R., Rath, T. M., and Manmatha, R. "Boosted decision trees for word recognition in handwritten document retrieval," in *Proc. of the 28th Annual Int'l ACM SIGIR Conf*, 2006, pp. 377-383.
31. Hu, J. Brown, M.K., Turin, W. HMM Based Online Handwriting Recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*. 1996.
32. Jones, K.S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11-20 1972.
33. Jones, K.S. and Willet, P. *Readings in Information Retrieval*, San Francisco: Morgan Kaufmann. 1997.
34. Kaufmann, G.; Bunke, H.; Hadorn, M. Lexicon Reduction in an HMM-Framework Based on Quantized Feature Vectors. *Proc. International Conference on Document Analysis and Recognition (ICDAR)* 97; Vol. 2 , 18-20, p.1097-1101. 1997.
35. Kim, G., Govindaraju, V. Bank Check Recognition using Cross Validation between Legal and Courtesy Amounts. *International Journal on Pattern Recognition and Artificial Intelligence*. 1996.
36. Kim, G., and Govindaraju, V.: A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. *IEEE Trans. PAMI* 19(4): 366-379. C1997.
37. Koerich, A.L., Sabourin, R., Suen, C.Y. Fast Two-Level HMM Decoding for Large Vocabulary Handwriting Recognition. *International Workshop on Frontiers in Handwriting Recognition*. 2004.

38. Koerich, A.L., Sabourin, R., Suen, C.Y. Large Vocabulary Off-Line Handwriting Recognition: A Survey. *Pattern Analysis and Applications*. 2003.
39. Larson, Hostetler, Edwards. *Calculus with Analytic Geometry*. Chapter 13 : Section 13.9. Fifth Edition. D.C. Heath and Company. USA. 1994.
40. Lopresti, D., and Zhou, J. Retrieval strategies for noisy text. In *Proceedings of Symposium on Document Analysis and Information Retrieval*, pages 255-270, 1996.
41. Luhn, H. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1: 309-317. 1957.
42. Madhvanath, S. The Holistic Paradigm in Handwritten Word Recognition and its Application to Large and Dynamic Lexicon Scenarios. University at Buffalo Computer Science and Engineering. Ph.D. Dissertation, 1997.
43. Madvanath, S., Kprasundar, V., Govindaraju, V. Syntactic Methodology of Pruning Large Lexicons in Cursive Script Recognition. *Journal. The Journal of the Pattern Recognition Society. Pattern Recognition* 34 (2001); Elsevier Science. 2001.
44. Marti, U.V., Bunke, H. Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition Systems. *World Scientific Series in Machine Perception and Artificial Intelligence Series*. 2001.
45. Milewski, R. and Govindaraju, V. Medical Word Recognition using a Computational Semantic Lexicon. *Eighth International Workshop on Frontiers in Handwriting Recognition*. Canada. 2002.
46. Milewski, R and Govindaraju, V. Handwriting Analysis of Pre-Hospital Care Reports. *IEEE Proceedings. Seventeenth IEEE Symposium on Computer-Based Medical Systems (CBMS)*. 2004.
47. Milewski, R and Govindaraju, V. Extraction of Handwritten Text from Carbon Copy Medical Forms. *Document Analysis Systems (DAS)*. Springer-Verlag. 2006.
48. Nakai, M., Akira, N., Shimodaira, H., Sagayama, S. Substroke Approach to HMM-Based On-Line Kanji Handwriting Recognition. *Sixth International Conference on Document Analysis and Recognition*. 2001.
49. National Library of Medicine. PubMed Stop List.
50. Oh, I.S., Suen, C.Y. Distance Features for Neural Network-Based Recognition of Handwritten Characters. *International Journal on Document Analysis and Recognition*. 1998.
51. Okuda, T. Tanaka, E. Kasai, T. A Method for the Correction of Garbled Words based on the Levenshtein Distance. *IEEE Transactions on Computers*, Col. C-25, No. 2. 1976.
52. Porter, M.F. An Algorithm for Suffix Stripping. *Program*, 14: 130-137. 1980.
53. Rath, T.M., and Manmatha, R. Features for word spotting in historical manuscripts. In *Proceedings of IEEE International Conference on Document Analysis and Recognition*, pages 218-222, 2003.
54. Rath, T. M., and Manmatha, R. Word spotting for historical documents," *IJDAR*, vol. 9, no. 2, 139-152, 2007.
55. Rath, T. M., and Manmatha, R. "Word image matching using dynamic time warping," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*, vol. 2, Madison, WI, June 18-20 2003, 521-527.
56. Rath, T.M., Manmatha, R., and Lavrenko, V. "A search engine for historical manuscript images," in *ACM SIGR*, 2004, pp. 369-376.
57. Rijsbergen, C.J. van, Robertson, S.E. and Porter, M.F. *New Models in Probabilistic Information Retrieval*. London: British Library. 1980.
58. Russell, G., Perrone, M.P., and Chee, Y.M. Handwritten document retrieval. In *Proceedings of International Workshop on Frontiers in Handwriting Recognition*, pages 233-238, 2002.
59. Salton, G. *Introduction to Modern Information Retrieval*. New York. McGraw-Hill. 1983.
60. Sinha, R.M.K., Prasada, B. Visual Text Recognition through Contextual Processing. *Pattern Recognition Vol 21, No 5*. pp.463-479. 1988

- 
61. Srihari, S.N., Hull, J.J., Choudhari, R. Integrating Diverse Knowledge Sources in Text Recognition. *ACM Transactions on Office Information Systems*. Vol.1 No.1 pp. 68-87. 1983.
  62. Suen, C.Y. N-gram Statistics for Natural Language Understanding and Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 1, No.2, pp 164-172. 1979.
  63. Taghva, K., Narkter, T., Borsack, J., Lumos, S., Condit, A., and Young, R. Evaluating text categorization in the presence of OCR errors. In *Proceedings of IS&T SPIE 2001 International Symposium on Electronic Imaging Science and Technology*, pages 68-74, 2001.
  64. Tan, C.L., Huang, W., Yu, Z., and Xu, Y. Imaged document text retrieval without OCR. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):838-844,2002.
  65. Vinciarelli, A. Application of Information Retrieval Techniques to Single Writer Documents. *Pattern Recognition Letters*, Vol. 26, no. 14-15, pp. 2262-2271, October 2005.
  66. Vinciarelli, A. Noisy Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, no. 12, pp. 1882-1295, 2005.
  67. Western Regional Emergency Medical Services. Bureau of Emergency Medical Services. New York State (NYS) Department of Health (DoH). Prehospital Care Report v4.
  68. Xue, H., and Govindaraju, V. Stochastic Models Combining Discrete Symbols and Continuous Attributes - Application in Handwriting Recognition. *Proceedings of 5th IAPR International Workshop on Document Analysis Systems*. pp. 70-81. 2002.
  69. Xue, H., and Govindaraju, V. On the Dependence of Handwritten Word Recognizers on Lexicons. *IEEE Trans. PAMI*, Vol. 24, No. 12, p. 1553-1564. 2002.
  70. Yates, B.R., and Ribeiro-Neto, B. *Modern Information Retrieval*, Addison-Wesley, 1999.
  71. Zimmermann, M. and Mao, J. Lexicon Reduction using Key Characters in Cursive Handwritten Words. *Pattern Recog. Letters*; Vol 20, p.1297-1304. 1999.
  72. Zobel, J., and Dart, P. Finding Approximate Matches in Large Lexicons, *Software Practice and Experience*, Vol 25, 3, 331-345, Mar 1995.